The Choice Is Yours! Researchers Assign Subject Metadata to Their Own Materials in Institutional Repositories

Maira Bundza
Western Michigan University, maira.bundza@wmich.edu

The Choice Is Yours! Researchers Assign Subject Metadata to Their Own Materials in Institutional Repositories

Maira Bundza, Western Michigan University

Abstract

The Digital Commons platform for institutional repositories provides a three-tiered taxonomy of academic disciplines for each item submitted to the repository. Since faculty and departmental administrators across campuses are encouraged to submit materials to the institutional repository themselves, they must also assign disciplines or subject categories for their own work. The expandable drop-down menu of about 1,000 categories is easy to use, and facilitates the growth of the institutional repository and access to the materials through the Internet.

The Digital Commons Institutional Repository Platform

Digital Commons is a platform for institutional repositories provided by Berkeley Electronic Press or *bepress* (http://digitalcommons.bepress.com/). The platform is being used by almost 300 institutions, mostly colleges and universities, but also includes consortia, health centers, associations, and governmental departments. Most are in the United States, but in 2013, nine were in Australia, six in Europe, and four in Asia.

These institutional repositories contain a wide variety of digital material. Much of the content is scholarly work done by faculty and students: published or unpublished articles and presentations, dissertations, theses, and other work including art and performances. The repositories contain publications such as journals, newsletters, books, pamphlets, brochures, and so on. They may also contain unique digitized collections. Some of the more unusual ones in Digital Commons are a restaurant menu collection from Johnson & Wales University in Providence, RI (http://scholarsarchive.jwu.edu/menus/) and a collection of letters between students at Bryant College in Smithfield, RI, and alumni serving their country during World War II (http://digitalcommons.bryant.edu/bryant_goes_to_war/). The largest collection currently on the platform is Digital Commons@University of Nebraska–Lincoln. In fall of 2013 they had over 65,000 items in their repository that had been downloaded over 19 million times. Paul Royster has worked hard to gather materials into their repository from all over campus. One of their most popular and unusual collections is from the Lester F. Larsen Tractor Test and Power Museum. The tractor test files are used worldwide and get 400,000 downloads annually, as Royster explained in a presentation about his repository at the 2011 Association of College Research Libraries Conference.[i] These examples give a sense of the scope of the collections—all of which need appropriate subject metadata.

The Digital Commons platform is based on a journal publishing program, so it provides a good platform for online journals—either new ones being created digitally or older publications with digitized back issues. About 550 journals were available in Digital Commons in fall 2013. The system provides an online submission process. The journal features include reviewing capabilities, and final acceptance or rejection of an article. As authors submit an article to a

journal, they add all the metadata in a form—information about themselves, as well as their submission. They provide their name, e-mail address, and affiliation. They have to provide a title for their submission, the type of submission it is, furnish keywords, an abstract, and choose a subject or discipline from the Digital Commons Three-Tiered Taxonomy.

The idea behind Digital Commons and other institutional repositories is that an institution would not need to hire an office full of employees to enter material into the repository, but instead people across the institution would be empowered to add materials themselves. Each institution has an administrator, but this person can give others administrative rights to sections of the repository. At Western Michigan University (WMU), the University Libraries have designated one person as the main administrator for their repository—ScholarWorks at WMU (http://scholarworks.wmich.edu/). A person in the Medieval Institute was given rights to add materials to the International Congress on Medieval Studies. The employees in the Provost's office that publish the newsletter *Prism* routinely put new issues into ScholarWorks. The Humanities Center coordinator adds materials to the Center's section of the repository. Eventually any faculty member will be able to upload material, and the designated administrator for his or her department will check the metadata and allow the material to be added to the site. The same submission process that is used for journal articles is also used throughout the repository.

SelectedWorks is a companion program to ScholarWorks that uses the same submission principles, but does not require the intervention of an administrator. SelectedWorks is a way for individual faculty members to create a professional page or online curriculum vitae, with full text of their work when copyright allows. This system uses a very similar form for each submission as in ScholarWorks and asks the author to choose from the same list of disciplines that will be described below. Currently SelectedWorks is more widely used than ScholarWorks by individual faculty members at WMU.

The Three-Tiered Taxonomy

The Digital Commons Three-Tiered Taxonomy of Academic Disciplines is a list of over 1,000 disciplines that can be added to each item in the repository. The disciplines are divided into three tiers. The topmost tier includes a broad division of disciplines:

Architecture
Arts and Humanities
Business
Education
Engineering
Law
Life Sciences
Medicine and Health Sciences
Physical Sciences and Mathematics
Social and Behavioral Sciences

The second tier breaks these broad disciplines into more specific disciplines, which may correspond to academic departments such as:

Chemistry
Computer Engineering
History
Marketing
Music
Psychology

The first two tiers function more as classifications than subject headings. It is usually the third tier that provides more detailed sub-disciplines or acts as a controlled subject vocabulary across all the repositories that use Digital Commons. An author or administrator can add one or more disciplines to an item from the Disciplines Widget, an expandable drop down menu (Figure 1).
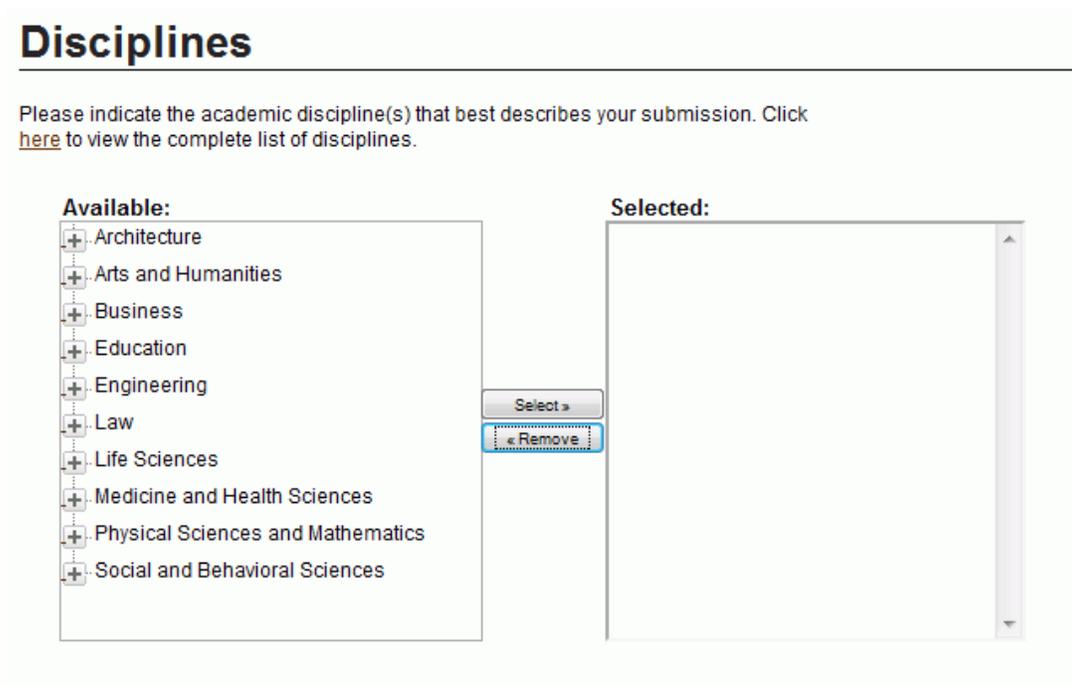

FIGURE 1 Top Tier Discipline Choices.

The term "Disciplines" can be modified to say "Subject Categories" or something else, as Western Michigan University has done for entering senior honors theses (see Figure 2).
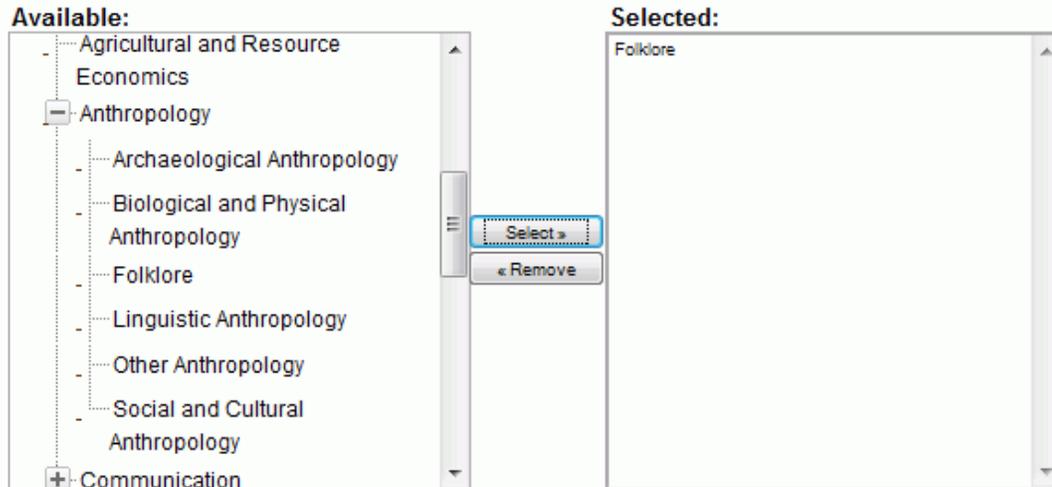
FIGURE 2 Third Tier of Disciplines for Anthropology.

In Figure 2, Anthropology under Social and Behavioral Sciences has been expanded to show six categories of Anthropology. In this case "Folklore" was chosen. If none of the categories fit, one can just choose "Anthropology" or "Other Anthropology."

Each Digital Commons repository has a Disciplines index on their site. To see how entries for Communication would look in a repository, see Figure 3.

FIGURE 3 Index of Disciplines on Repository Site for Browsing.

Note that there are 712 items in the repository under Social and Behavioral Sciences, 52 of which are under Communication. Only 11 of the Communication entries have been given third tier categories, so the rest were simply put under the broader discipline of Communication.

Instead of working with the five-volume *Library of Congress Subject Headings* list that needs cataloging skills to use effectively, this simple 1,000-plus discipline list can be used easily by authors and departmental administrators of repository pages. Authors know the subject of their work the best; know to which category they belong, and understand how their discipline is divided into more specific topics. They may not find their specialization, but they can always choose a broader discipline and can add keywords to facilitate searching by search engines.

Above the expandable drop-down menu widget there is a link to a 25- page PDF containing the full list that can be printed, skimmed, or searched online. For example, WMU has published the journal *Reading Horizons* for over 50 years. Some of the articles talk about early childhood education or other educational topics, but most of the articles are specifically about reading. Searching for the word "reading" locates "Reading and Language" under "Arts and Humanities." To indicate the educational aspect of the articles in this journal, one may choose to add "Education" as another discipline.

Kenneth Gleason from Digital Commons maintains the list and explained that his company decided to build the taxonomy as a way to connect the repositories together in a meaningful fashion. They saw the potential of uniting all of the Digital Commons repositories through common subject headings. They added the Disciplines Widget (see Figures 1 and 2) as a means of relating articles across multiple repositories through these common subject headings. They are building new and exciting navigation tools across repositories for their users based on the power of this connection.

The introduction to the list explains that Digital Commons developed this list with the help of the University of California's California Digital Library. They used the following sources:
- Taxonomy of Research Doctoral Programs from the National Academies, http://sites.nationalacademies.org/PGA/Resdoc/PGA_044478
- Classification of Instructional Programs, 2000 edition, from the National Center for Educational Statistics (NCES), http://nces.ed.gov/pubs2002/cip2000
- Medical Subject Headings (MeSH) from the National Library of Medicine, http://www.nlm.nih.gov/mesh/2007/MeSHtree.html
- Current Index to Legal Periodicals (CILP) and FindLaw
- Several Business categories come from Cabell's
- The University of California's list of departments and programs

The Digital Commons Three-Tiered Taxonomy of Academic Disciplines provides a controlled vocabulary system that can be used for most academic research.[ii] To maintain relevance, Digital Commons asks participating institutions to send in suggestions for improving the list. They evaluate the suggestions, merging similar disciplines and adding new ones, and include the updates in each quarterly release. As the range of institutions and repositories expands, the scope of the list may also expand. One can already see the effects of the participating law and medical schools on the list.

Categories of the Disciplines

Architecture is the smallest discipline and has only two tiers with 10 subcategories such as Environmental Design and Landscape Architecture. Architecture also appears in Arts and Humanities, under the History of Art, Architecture, and Archaeology.

Arts and Humanities is one of the largest sections, as it includes Art and Design, Classics, History, Languages and Literatures, Music, Philosophy, Religion, Theatre, and the studies of various cultures. As Digital Commons spreads across the world, the editors may find themselves expanding this list.

Business and Education have fairly short lists of subcategories, with only a few areas expanded to the third tier (e.g., Teacher Education and Professional Development). The use of these categories may depend on the presence of a business school or a college of education, although any material about a college or university itself may also be listed under Education.

Engineering has an extensive list with categories under Aerospace, Biomedical, Chemical, Civil and Environmental, Computer, Electrical, Materials Science, Mechanical and other engineering disciplines.

The Law section was developed based on the needs of the 60 participating law schools that have put 150 law journals in Digital Commons. This is the only other section besides Architecture that is not three-tiered. Under the broad discipline of Law, there are over 100 law related categories such as Cyberspace Law, Immigration Law, National Security, Tax Law, and so on.

Life Sciences includes Animal Sciences, numerous divisions of Biology, Food Science, Genetics, Immunology, Kinesiology, Neuroscience, Nutrition, Pharmacology, Physiology, and Plant Sciences, just to name a few of the main divisions. Some of these disciplines may overlap with Medicine and Health Sciences.

The lengthiest list, with more than 200 categories, is for Medicine and Health Sciences. Only nine medical schools are listed as Digital Commons repositories, but many universities may have a medical program or a college of health sciences. This list has been based on the Medical Subject Headings (MeSH) Tree Structure, but has expanded to include Dentistry, Medical Specialties, Nursing, Public Health, Rehabilitation, and Veterinary Medicine, to name just a few of the other disciplines.

Physical Sciences and Mathematics also have a long list with many subcategories for Applied Mathematics, Astrophysics, Chemistry, Computer Sciences, Earth Sciences, Environmental Sciences, Mathematics, Oceanography, Physics, and Statistics and Probability. Computer Sciences is found both here and under Engineering. At WMU the computer science department moved from the College of Arts and Sciences to the College of Engineering and Applied Sciences, so this seeming duplication may help categorize a document according to the institutional structure.

Last, but not least with almost 150 disciplines, are the Social and Behavioral Sciences, which include among others, Anthropology, Communication, Economics, Geography, International and Area Studies, Legal Studies, Linguistics, Political Science, Psychology, Public Administration, and Sociology.

Practical Use of the Taxonomy

One of the most diverse projects in the WMU repository is the inclusion of honors theses. To graduate from the Lee Honors College at Western Michigan University, each student must submit an honors thesis, usually in the form of a written paper. Honors students come from all seven academic colleges and many different departments across campus. The Honors College asked the Library to digitize over 2,200 theses from their files going back to 1963. When a box of theses has been digitized, the theses are entered into the ScholarWorks repository. The full text of the theses is made available *only* if the student has signed a permission form.

The person entering the theses into ScholarWorks usually assigns one to three disciplines to each thesis, as many of the honors students have a major and minor concentration or even a double major, and have tried to integrate both subject areas into their thesis. The person entering the data has found the taxonomy easy to use in most cases, and finds it very helpful to be able to search the whole list online, if the discipline she is looking for is not where she expected it to be. Some areas that are not in her fields of expertise are confusing, such as the Medicine and Health Sciences. Another challenge is when some things appear in more than one place (e.g., Gender and Sexuality is found under Social and Behavioral Sciences: Sociology as well as under Arts and Humanities: Feminist, Gender and Sexuality Studies). This may be resolved with continued improvements to the list or it may usefully reflect that these topics are covered by different departments at various universities.

The Library and Information Science field is found under both Engineering and Social & Behavioral Sciences. There is definitely a technical aspect to our field, especially in the world of digital information that belongs under engineering. Much of what is covered in Library and Information Science classes probably falls more logically under social sciences.

Comparison with Other Repositories

There are other platforms used for institutional repositories, but only one was found to provide a controlled vocabulary for subject metadata similar to the Digital Commons Three-Tiered Taxonomy. Content DM and Fedora are repository platforms that allow use of various subject metadata schemes. The platforms below were looked at more closely.

DSpace (http://www.dspace.org/) is a well-known open access software developed for repositories by the Massachusetts Institute of Technology and Hewlett Packard Labs. DSpace is customizable and currently has around 1,300 participating institutions according to their Web site. Chapman, Reynolds, and Shreeves looked at metadata for three DSpace institutional repositories and found that no controlled vocabulary was used consistently across those institutions.[iii] Each collection within these repositories was allowed to use metadata appropriate

to the discipline, including the choice of subject metadata. Some chose to just use keywords, while others used a thesaurus or controlled vocabulary appropriate for their discipline.

EPrints 3 (http://wiki.eprints.org/) is a generic repository building software developed by the University of Southampton in England. Its documentation manual explains how each institution can set up its own set of subjects under the "Organisation Hierarchy" section.

arXiv (http://arxiv.org/) is a repository for research articles in computer science, condensed matter, mathematics, nonlinear sciences, physics, quantitative biology, and statistics maintained by Cornell University Library. Registered authors submit articles to this archive and may choose one or more subjects for their article. The primary categories were listed above, and each category has a limited set of more granular subjects for about 120 total subject classes. In this aspect it is most similar to the Digital Commons taxonomy.

OAI Compliance and the Semantic Web

All institutional repositories examined in this article comply with the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). Gleason explained in an e-mail: "Currently, bepress Digital Commons is not involved in the semantic web; but it's something we're looking at doing for the future. . . . Currently, our focus is on discoverability of objects in the repository via harvesters, search engines, and end users."[iv] They keep their Discipline names in line with the sources they have used (see the list under The Three-Tiered Taxonomy section of this article), so identifiers could be assigned later. They have their own Author Disambiguation tool and are following the work of Open Researcher & Contributor ID (ORCID). Hopefully this will connect to other name authority files in the future, so works by individuals can be made available across print and digital collections.

Conclusion

The Digital Commons Three-Tiered Taxonomy of Academic Disciplines provides a controlled vocabulary for institutional repositories that is simple enough to be used by researchers and administrators adding materials to the repository. This provides a uniform way of organizing materials across the institutional repositories that are using the Digital Commons platform, allows searching across repositories, and helps optimize research for discovery by search engines.

3. John W. Chapman, David Reynolds, and Sarah A. Shreeves, "Repository Metadata: Approaches and Challenges," *Cataloging & Classification Quarterly* 47, nos. 3–4 (2009): 309–325. http://dx.doi.org/10.1080/01639370902735020 (accessed October 3, 2013).
4. Kenneth Gleason, Product Manager at Digital Commons. Personal interview, December 14, 2011; e-mails, February 23 and February 29, 2012.
Downloaded by

---

[i] Paul Royster, "ACRL Tailgate 2011: 1.) IR Content from 'Off the Beaten Track' and 2.) Books and Such," *Library Conference Presentations and Speeches*, http://digitalcommons.unl.edu/library_talks/63 (accessed October 3, 2013).

[ii] Berkeley Electronic Press, "Disciplines: Introduction to Digital Commons Three-Tiered Taxonomy of Academic Disciplines," *Digital Commons Reference Material and User Guides* (2012), http://digitalcommons.bepress.com/reference/9 (accessed Sept. 10, 2012).

[iii] John W. Chapman, David Reynolds, and Sarah A. Shreeves, "Repository Metadata: Approaches and Challenges," *Cataloging & Classification Quarterly* 47, nos. 3-4 (2009): 309-325. http://dx.doi.org/10.1080/01639370902735020 (accessed Feb. 8, 2012).

[iv] Kenneth Gleason, Product Manager at Digital Commons. Personal interview, Dec. 14, 2011; emails, Feb. 23 and Feb. 29, 2012.